Unit 7:

* Still doing inference (tests and intervals)
* Will be doing inference to see if 2 variables are related
* Ch. 26 = ***categorical*** variables
* Ch. 27 = ***quantitative*** variables

CH. 26: CHI- SQUARE TESTS

There are 2 tests (no intervals):

1- Goodness of Fit
2- Homogeneity/Independence

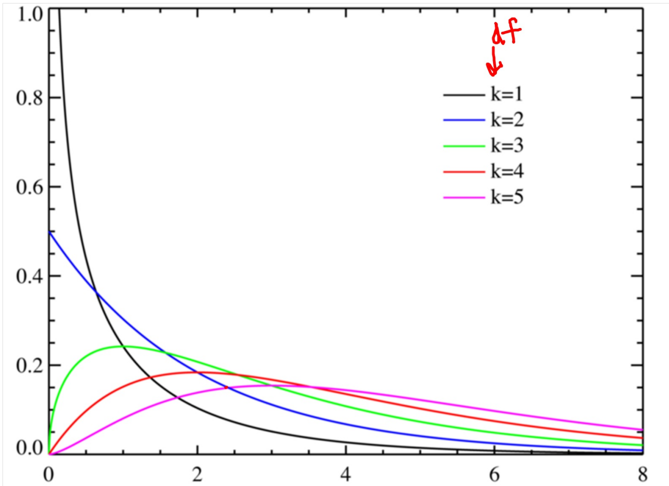
All use the Chi-Square variable for the test statistic: χ^2

All have the same formula for the test statistic:

$$\chi^2 = \sum \frac{(\text{observed count -- expected count})^2}{\text{expected count}}$$

Chi-Square distribution:

* Uses df (there is more than one distribution)
* Skewed right
* Picture: page A-77 in book
  or Table C on formula sheet
* Less skewed right as df increases:
  * if df = ∞, then χ^2 = normal



Goodness of Fit

these examples are not in the notes... just read thru them

EXAMPLE1:

I want to test to see if a die is fair. So I roll the die 60 times and find the following distribution.

observed:	expected:
1 = 13	10
2 = 8	10
3 = 18	10
4 = 8	10
5 = 7	10
6 = 6	10

Does the die seem fair?

NO! there are a lot of 3's!

But is this ONE category being this different enough to say the WHOLE die isn't fair??

EXAMPLE 2:I want to see if the distribution of colors of skittles is the same as what is claimed by the company. The company claims that the distribution is:

RED	GREEN	ORANGE	YELLOW	PURPLE	
30%	10%	20%	15%	25%	(expected %)
60	20	40	30	50	(expected # with 200 skittles)

I take a sample of 200 skittles and get the following results:

RED	GREEN	ORANGE	YELLOW	PURPLE	
33	12	31	10	14	(observed #)

Do the observed and expected seem to match up??

Ch-Square Goodness of Fit test

Used to test whether an observed distribution fits/matches an expected distribution

Hypotheses: (written sentences, always in context!)

Ho: Observed distribution of _____ fits the expected distribution of _______

Ha: Observed distribution of ______doesn't fit the expected distrib. of _______

Conditions:

1) Categorical data — 1) make a statement that the data is categorical
2) SRS
3) all expected counts are ≥ 5 — 3) show all expected counts and then state that they are all above 5.

conditions met --> χ^2 distribution --> χ^2 GOF test

Mechanics:

Test statistic:

write the formula — show the first 2 calculations

$$\chi^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}} = \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \ldots = \_\_\_$$

fill in the total (the test stat)

P-value:

$P(\chi^2 > \underline{\text{test statistic}}) = \chi^2\text{cdf(test stat, E99, df)}$

same place as normcdf

df = # categories (or outcomes) - 1

EX: Dice rolls: 6 outcomes/categories, df = 5

Conclusion:

same 2 sentences:

- reject/fail to reject Ho....
- Sufficient/Insufficient evidence... (for the Ha)

EXAMPLE: Testing to see if a die is loaded

What is the expected distribution (in %) for any fair die?

Number	1	2	3	4	5	6
%	1/6	1/6	1/6	1/6	1/6	1/6

I roll the die 120 times. What is my expected distribution (in #'s)? (these are the expected counts)

Number	1	2	3	4	5	6
Frequency	20	20	20	20	20	20

(these are the observed counts)

I roll the die 120 times. Below is the actual distribution:

Number	1	2	3	4	5	6
Frequency	18	18	17	18	17	32

Hypotheses:

Ho: Dice rolls are uniformly distributed

Ha: Dice rolls are not uniformly distributed

Conditions:

1) Categorical data — 1) rolls of die are categ. var.
2) SRS — 2) assumed representative
3) all expected counts are ≥ 5 — 3) all exp. counts = 20 ≥ 5

conditions met --> χ^2 distribution --> χ^2 GOF test

Test Statistic:

$$\chi^2 = \sum \frac{(obs - exp)^2}{exp} = \frac{(18-20)^2}{20} + \frac{(18-20)^2}{20} + \ldots = 8.7$$

$$= \frac{(18-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(32-20)^2}{20} = 8.7$$

DF = 5

P-Value: $P(\chi^2 > 8.7) = 0.1216$

χ^2cdf(8.7, E99, 5)

Conclusion:

We fail to reject Ho b/c p-value of 0.1216 > alpha = 0.05.

We have insufficient evidence that the dice rolls are not uniformly distributed. We have insufficient evidence that that die is unfair.

Example #2:

Does color impact the chance that a car is stolen? Suppose it is known that of all cars 15% are white, 15% are blue, 35% are red, 30% are black, and 5% are other colors. A random sample of 830 stolen cars is taken and their colors are noted below. Test to see if the cars stolen match the distribution of all cars.

Color	**White**	**Blue**	**Black**	**Red**	**Other**
Observed	140	100	230	270	90

(these are the observed counts)

We need to find the expected counts. Take sample size (n) and multiply by each of the given expected percents.

Color	**White**	**Blue**	**Black**	**Red**	**Other**
EXPECTED	124.5	124.5	290.5	249	41.5

Hypotheses:

Ho: The distribution of the colors of stolen cars matches the distribution of all cars.

Ha: The distribution of the colors of stolen cars doesn't match the distribution of all cars.

Conditions:

1) Categorical data	1) color of car is a categorical variable
2) SRS	2) stated random
3) all expected counts are ≥ 5	3) Expected counts all ≥ 5 (124.5, 124.5, 290.5, 249, 41.5)

conditions met --> χ^2 distribution --> χ^2 GOF test

Mechanics:

$$\chi^2 = \sum \frac{(obs - \exp)^2}{\exp} = \frac{(140-124.5)^2}{124.5} + \ldots + \frac{(90-41.5)^2}{41.5} = 66.33$$

df = 4

$$P-Value = P\left(\chi^2 > 66.33\right) = 1.35 \times 10^{-13}$$

χ^2cdf(66.33, E99, 4)

Conclusion:

We reject Ho b/c p-value < alpha = 0.05

We have sufficient evidence that the distribution of the colors of stolen cars does not fit the distribution of all cars.

Calculator:

L1 = observed values
L2 = expected values
L3 = (obs - exp)^2/exp = (L1 -L2)^2/L2

sum L3: 2ND LIST --> MATH --> #5: sum --> ENTER

sum(L3) = χ^2 test statistic

P-value: χ^2cdf(test stat, E99 df)

Try the example problems in the notes

#1 -- 4

1) A professor of education classes at Virginia Tech wants to look at what types of education the VT students are choosing. From previous studies, the types of education have been known to have the following distribution: 25% physical education, 15% math education, 15% science education, 5% art education, 20% special education, 10% history education, 5% foreign language education, and 5% other. He takes a random sample of 154 education majors and finds the following results: 40 phys ed, 20 math, 10 foreign language, 30 special ed, 15 history, 20 science, 10 art, and 9 other. Has the distribution of education majors changed? Run a full test of significance.

H_0: The distribution of majors matches the historic distribution.
H_A: The distribution of majors is not the same as the historic distribution.

Conditions:
1) Categorical Data: choice of major is categorical

2) SRS: Stated as a random sample

3) All Exp Cell counts $\geq$ 5:

Major	PE	ME	SE	AE	SpE	HE	FLE	O
Expected	38.5	23.1	23.1	7.7	30.8	15.4	7.7	7.7

All expected cell counts are greater than 5.

Conditions met → Chi-Square goodness-of-fit test

Mechanics: df = $k - 1 = 7$

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

$$\chi_7^2 = \frac{(40-38.5)^2}{38.5} + \ldots + \frac{(9-7.7)^2}{7.7} = 2.515$$

$$P - Value = P\left(\chi_7^2 > 2.515\right) = 0.926$$

Since the P-Value is greater than alpha (0.926 > 0.05) we fail to reject the null hypothesis. There is not enough statistically significant evidence that the distribution of ed. majors is different than it has historically been.

2) A grocery store manager wishes to determine whether a certain product will sell equally well in any of five locations the store. Five displays are set up, and the resulting numbers of the product sold are 43, 29, 52, 34, and 48. Is there enough evidence that the location makes a difference (are the locations **equally** as popular, or not)? Test at both the 5% and 10% significance levels.

H_0: The distribution of sales is the same for all 5 locations.

H_A: The distribution of sales is not the same for all 5 locations.

Conditions:

1. Counted Data: The data is counts of sales in different locations.
2. Randomization: Assume Representative
3. Expected Cell Frequency:

Location	1	2	3	4	5
Expected	41.2	41.2	41.2	41.2	41.2

All expected cell counts are greater than 5.

All conditions have been met for a Chi-Square goodness-of-fit test.

Mechanics:

df = $n - 1 = 4$ $\alpha = 0.05$ $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$

$$\chi_4^2 = \frac{(43 - 41.2)^2}{41.2} + \ldots + \frac{(48 - 41.2)^2}{41.2} = 8.903$$

$$P - Value = P\left(\chi_4^2 > 8.903\right) = 0.0636$$

Since the P-Value is greater than alpha (0.0636 > 0.05) we fail to reject the null hypothesis.

There is not enough statistically significant evidence that the distribution of sales is different between the 5 locations

3) A program for generating random numbers on a computer is to be tested. The program is instructed to generate 100 single-digit integers between 0 and 9. The frequencies observed are 11, 8, 7, 7, 10, 10, 8, 11, 14, and 14. Is there sufficient reason to believe that the integers are not being generated uniformly?

#3:

Ho: the distribution of numbers generated is uniform

Ha: the distribution of numbers generated is not uniform

Conditions:

1) categorical	1) the outcomes of the generator are categorical
2) SRS	2) assumed representative
3) expected counts ≥ 5	3) all exp. = $10 \geq 5$

conditions met --> χ^2 distribution --> χ^2 GOF test

$$\chi^2 = \Sigma\frac{(obs - exp)^2}{exp} = \frac{(11 - 10)^2}{10} + \frac{(8 - 10)^2}{10} + \ldots = 6$$

$P(\chi^2 > 6 | df = 9) = 0.7399$

We fail to reject Ho b/c p-value of 0.7399 > alpha = 0.05. We have insufficient evidence that the random numbers are not generated uniformly. The generator seems to be fair.

4) Portable personal computers, or "laptops," represent a fast-growing segment of the PC market. According to Market Intelligence Research company, the use of laptops can be classified in the following user segments ("Laptop's Three Musts," 1988): Business-professional (69%), Government (21%), Education (7%), and Home (3%). 200 laptop owners were surveyed this year, and the user segments were tabulated as follows: Business-professional (102), Government (32), Education (22), and Home (44). Do the data provide sufficient evidence to indicate that the figures given in 1988 are not accurate today?

#4:
Ho: the distribution of use of laptops fits the given percents
Ha: the distribution of use of laptops doesn't fit the given percents

Conditions:

1) categorical — 1) the outcomes are categorical
2) SRS — 2) assumed
3) expected counts ≥ 5 — 3) all expected counts ≥ 5 (138, 42, 14, 6)

conditions met --> χ^2 distribution --> χ^2 GOF test

$$\chi^2 = \Sigma\frac{(obs - exp)^2}{exp} = \frac{(102 - 138)^2}{138} + \frac{(32 - 42)^2}{42} + \ldots = 257.01$$

$P(\chi^2 > 257.01 | df = 3) = 1.993 \times 10^{-55}$

We reject Ho b/c p-value of $1.993 \times 10^{-55} <$ alpha $= 0.05$. We have sufficient evidence that laptop use does not fit the given percentages

HW: p. 643 #4, 5, 7

Chi-Square test #2: Homogeneity/ Independence

* Uses 2-way tables (2 variables)
* Compares 2 variables that are measured from the same sample

Example: Gender vs. college location

	In-State	Out-of-State
MALE	25	36
FEMALE	37	41

Want to see if there is an association between the 2 variables

Homogeneity: Compares the distribution of several groups for the same categorical variable.

Is the distribution of grades the same throughout the years?

Is the distribution of gender the same throughout the rank in the military?

Are there differences in the distributions of type of car between students and staff?

Is the distribution of political party different between gender?

PAGE 646 #28

Independence: Examines counts from a single group to test for an association between two categorical variables

Look for: association, affect, independence, relationship

Examples:
Is there an association between grades and years?

Does gender affect someones rank in the military?

Are type of car and type of driver independent?

Is there a relationship between political party and gender?

PAGE 646 #29

Hypotheses: Depends on the type of test.
Write the hypotheses so that they match the type of test

Ho: $\overbrace{\text{equal = no difference}}^{H}$ = $\overbrace{\text{independent = no association}}^{I}$

Ha: not equal = there IS a difference = dependent = assoc.

Conditions: Same 3 as the GOF test
1- Categorical data
2- SRS
3- All expected cell counts ≥ 5

conditions met --> χ^2 distribution -->
χ^2 test for Homogeneity/Independence

Mechanics:
Test statistic: same as before

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \_\_\_\_\_\_ + \_\_\_\_\_\_ + \ldots =$$

$$\text{Expected cell count} = \frac{(\text{row total})(\text{column total})}{n}$$

Example: Expected counts for gender vs. college location:

obs:

	In-State	Out-of-State	
MALE	25	36	61
FEMALE	37	41	78
	62	77	139

Exp:

	In-State	Out-of-State
MALE	$\frac{61 \times 62}{139} = 27.21$	33.79
FEMALE	$\frac{78 \times 62}{139} = 34.8$	43.21

P-value:

P(χ^2 > test statistic) = χ^2cdf(test stat, E99, df)

df = (r - 1)(c - 1)   #rows   #columns

Conclusion
same 2 sentences.
* reject/fail to reject Ho....
* Sufficient/insufficient evidence that (re-copy Ha)

Calculator
- put observed in Matrix A
- Run χ^2 test
- expected values are now in Matrix B

Example: A local college admissions officer wants to know if student choices in majors have changed through the years. He selects a random sample of freshman from 1995, 2000, and 2005 and check's their choice of major.

	1995	2000	2005
Buisness	45	79	63
Education	25	34	43
Humanities/Arts	37	46	23
Science	26	62	72
Undecided	37	29	49

Test to see whether the majors are distributed the same each year.

Hypotheses:
Ho: The choice of majors have the same distribution throughout the years.

Ha: The choice of majors do not have the same distribution throughout the years.

Conditions:
1) Categorial data   ① major + 3 years are categ.
2) SRS   ② stated random
3) all expected cell counts ≥ 5 ③ lowest exp. count=25.881≥5

conditions met --> χ^2 distribution -->
χ^2 test for Homogeneity

Test Statistic:

$\chi^2 = \sum \frac{(obs-exp)^2}{exp} = \frac{(45-47.448)^2}{47.448} + \frac{(79-69.776)^2}{69.776} + \ldots =$

$df = 8$

30.222

P-Value: $P(\chi^2 > 30.222) = 1.9307 \times 10^{-4}$

- Reject Ho....
- Suff. evid.... (recopy Ha)

Example: Smoking vs. Socio-Economic Status (SES). Below is a table of a random sample of 356 PA residents, giving the Smoking status versus Socio-Economic status (SES). Test whether there is an association between the two variables. Does there appear to be a relationship between smoking and socioeconomic status?

	SES		
Smoking	High	Middle	Low
Current	51	22	43
Former	92	21	28
Never	68	9	22

Hypotheses:

Ho: There is no association between smoking and socioeconomic status (independent) (no relationship)

Ha: There is an association between smoking and socioeconomic status (dependent) (relationship)

Conditions:

1) Categorical data	1) smoking & SES are categ.
2) SRS	2) stated
3) all expected counts ≥ 5	3) lowest exp count = 14.461 ≥ 5

conditions met --> χ^2 distribution -->
χ^2 test for Independence

Test Statistic:

$\chi^2 = \Sigma \frac{(obs - exp)^2}{exp} = \frac{(51 - 68.753)^2}{68.753} + \frac{(22 - 16.944)^2}{16.944} + \ldots =$

$= 18.51$

P-Value:

$P(\chi^2 > 18.51 \mid df = 4) = 9.808 \times 10^{-4}$

Conclusion:

We reject Ho b/c p-value of 9.808×10^{-4} < alpha = 0.05.

We have sufficient evidence that there is a relationship between smoking and SES.

Worksheet in notes #1:

1) The manager of an assembly process wants to determine whether the number of defective products manufactured depends on the day of the week the articles are produced. Using the data below, is there sufficient evidence to determine if the distribution of defective products is the same throughout the work week?

Day	Mon.	Tue.	Wed.	Thur.	Fri.
Nondef.	85	90	95	95	90
Defective	15	10	5	5	10

Hypotheses:

Ho: The distribution of defective products is the same throughout the days of the week

Ha: The distribution of defective products is NOT the same throughout the days of the week

Conditions:

1) Categorical data	1) defectiveness and days of week are categ.
2) SRS	2) assumed representative
3) all expected counts ≥ 5	3) lowest exp count = $9 \geq 5$

conditions met --> χ^2 distribution --> χ^2 test for Homogeneity

Test Statistic:

$$\chi^2 = \Sigma\frac{(obs - exp)^2}{exp} = \frac{(85 - 91)^2}{91} + \frac{(90 - 91)^2}{91} + \ldots =$$

$$= 8.547$$

P-Value:

$P(\chi^2 > 8.547 | df = 4) = 0.0735$

Conclusion:

We fail to reject Ho b/c p-value of 0.0735 > alpha = 0.05. We have insufficient evidence that the distribution of defective products is not the same throughout the days of the week.

Worksheet in the notes, #2:

2) The following table is from the July 1993 publication of *Vital and Health Statistics* from the Centers for Disease Control and Prevention/National Center for Health Statistics. The individuals in the following table have only one of the three indicated irritations. Determine if the type of irritation is independent of the age group using a 0.05 level of significance.

Irritation	18-29	30-44	45-64	65+
Eye	440	567	349	59
Nose	924	1311	794	102
Throat	253	311	157	19

HW:

p. 643 #8, 10, 13, 14, 24

- Read each problem
- Determine which type of test
(GOF, Homogeneity, Independence)
- Write the hypotheses for the test (Ho only)

DO problems #23, 29